Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 90 (2016) 119 – 124

Procedia Computer Science

International Conference On Medical Imaging Understanding and Analysis 2016, MIUA 2016, 6-8 July 2016, Loughborough, UK

Persistent Homology for Fast Tumor Segmentation in Whole Slide Histology Images

Talha Qaiser[a], Korsuk Sirinukunwattana[a], Kazuaki Nakane[b], Yee-Wah Tsang[c], David Epstein[d], Nasir Rajpoot[a,*]

[a]*Department of Computer Science, University of Warwick, Coventry, CV4 7AL, United Kingdom*
[b]*Osaka University, Suita, 565-0871, Japan*
[c]*University Hospital of Coventry and Warwickshire, Coventry, CV2 2DX, United Kingdom*
[d]*Mathematics Institute, University of Warwick, Coventry, CV4 7AL, United Kingdom*

Abstract

Automated tumor segmentation in Hematoxylin & Eosin stained histology images is an essential step towards a computer-aided diagnosis system. In this work we propose a novel tumor segmentation approach for a histology whole-slide image (WSI) by exploring the degree of connectivity among nuclei using the novel idea of *persistent homology profiles*. Our approach is based on 3 steps: 1) selection of exemplar patches from the training dataset using convolutional neural networks (CNNs); 2) construction of persistent homology profiles based on topological features; 3) classification using variant of k-nearest neighbors (k-NN). Extensive experimental results favor our algorithm over a conventional CNN.

Peer-review under responsibility of the Organizing Committee of MIUA 2016

Keywords: Digital Pathology; Tumor Segmentation; Histology Image Analysis; Persistent Homology; Colorectal Cancer;

1. Introduction

Colorectal cancer is the second most commonly diagnosed cancer in females and the third most in males, with an estimated 1.4 million cases and 693,000 deaths occurring in 2012[1]. Identification of tumor-rich areas from histopathology colon images is one of the primary tasks for a computer-aided diagnosis system. Automated tumor segmentation methods are therefore highly desirable and have received much research effort during recent years.

Recently published techniques for tumor segmentation mostly rely on morphological appearance and local texture features of tissue components[2]. However, the high variability of features in tumor regions and tissue samples pose the risk of over-emphasizing the spatial properties of a particular dataset; these methods do not produce the desired results. Some methods focus on nucleus segmentation or sub-cellular components of tissues[3,4]. However, cell segmentation is non-trivial due to the atypical characteristics and heterogeneous appearance and, often, clumped structure of cancerous cells, so these methods are also likely to fail. For these reasons, detection of tumor regions is a highly challenging problem.

* Corresponding author
E-mail address: N.M.Rajpoot@warwick.ac.uk

Peer-review under responsibility of the Organizing Committee of MIUA 2016
doi:10.1016/j.procs.2016.07.033

In tumor regions nuclei have different atypical characteristics, with non-uniform chromatin texture and irregularity in their shape and size. In tumor areas nuclei clump together, filling the inter-cellular regions, and the structure of individual nuclei becomes more difficult to discern. In contrast, nuclei remain relatively distinct in normal regions and maintain their structure and appearance (Fig 1). The arrangement of nuclear structures is a significant feature for tumor classification. In this paper, we propose a new approach to tumor segmentation in WSIs by characterizing the connectivity between cells, using persistent homology. Our approach works in the following ways : 1) a method for constructing persistent homology profiles based on topological features; 2) an algorithm for selection of exemplar patches from the training dataset based on highly activated nodes of a CNN; and 3) a variant of k-NN classifier.

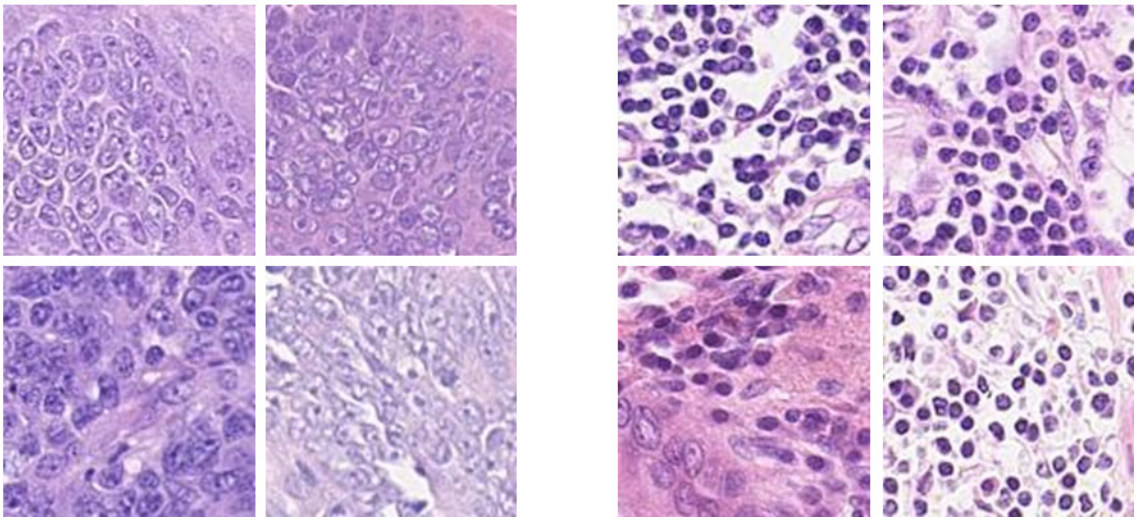


Fig. 1: (a) Tumor patches, exhibiting non-uniformity in chromatin texture and atypical characteristics. (b) Non-tumor patches showing the homogeneous structure of normal nuclei.

2. Proposed Algorithm

Given a colorectal WSI, we first divide it into patches. The problem is to then classify each patch as tumor or non-tumor. We approach this problem by exploring the connectivity between cells. We associate with each patch two statistical distributions, which we call *persistent homology profiles*, computed using certain topological invariants, as will be explained below. For simplicity, we will explain our method using a single persistent homology profile for each patch. In that case, given two patches, we have the persistent homology profile for each, and the symmetrised Kullback-Leibler divergence (KLD) between these two distributions gives a numerical estimate for how far the patches are from each other. An input patch Q is classified by a k-NN classifier, based on KLD distances between the persistent homology profiles of representative patches, denoted by P (Fig 2). We decide on our representative patches by training a CNN, and then selecting, separately for the tumor class and the non-tumor class, patches whose activation during the training is large. The essence of this approach is to use the subset of highly activated patches from convolutional layers as exemplars, rather than using the set of all the patches in the training dataset.

2.1. Persistent Homology

2.1.1. Overview.

Persistent homology is a fairly recent concept, surveyed in[5] and[6]. Let M denotes a graylevel image of size mxn, where gray intensities are integer values between 0 and 255 and B be an $m \times n$ closed rectangle. For each value of the threshold t, let $B(t) \subset B$ be the union of pixels with intensity less than or equal to t. We want $B(t)$ to be a closed subspace of B, and so, with each pixel in $B(t)$, we include its four corners and four sides. In effect, we binarize the intensity values in M, replacing any value less than or equal to t by 0 and replacing other values by 1. The resulting matrix $M(t)$ gives us a black and white image. The notation is supposed to remind us that $B(t)$ is the union of black pixels. We have $B(0) \subseteq B(1) \subseteq \cdots \subseteq B(255) = B$, so that B is a filtered space, persistent homology's essential ingredient. We will work with the zeroth Betti number $\beta_0 : [0, 255] \cap \mathbb{Z} \to \mathbb{Z}$ which maps a threshold value t to the

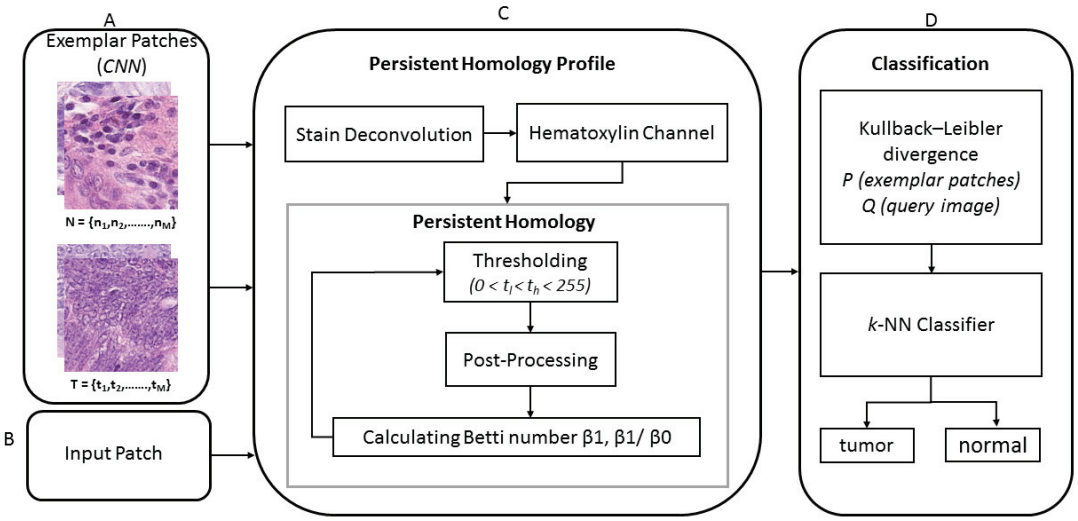


Fig. 2: (A) CNN based exemplar patches selected from the training dataset. (B) Input patch for testing. (C) Algorithm for persistent homology profiles. (D) Determine KLD between testing and exemplar patches and classify through k-NN.

number of components of $B(t)$: this can be quickly computed from the matrix M. We are also interested in the first Betti number $\beta_1 : [0, 255] \cap \mathbb{Z} \to \mathbb{Z}$, giving the number of independent circuits in $B(t)$, or, equivalently, since we are working in the plane, the number of holes in $B(t)$. To compute $\beta_1(t)$, we first find the white components, that is the components of the complement $W(t) = B \setminus B(t)$ of $B(t)$. As t increases, new components appear in $B(t)$ and old components increase in size and amalgamate. Also, as t increases, and black pixels replace white pixels, black regions may coalesce to surround new white holes, or white holes may split into several white holes or disappear completely. Each of the two functions β_0 and β_1 sometimes increases and sometimes decreases as t varies (see Fig 3), depending on the details of the image matrix M.

2.1.2. Persistent Homology Profiles (PHP)

A block diagram showing the computation of a persistent homology based profile is shown in Fig 2C. Histology images stained on different occasions often vary considerably in color. To overcome this problem, we first carry out stain normalization [7] to obtain consistency. For each threshold t, we convert the normalized image into a binary image, prior to computing persistent homology. In tumor regions, nuclei have atypical characteristics and lie relatively close to each other. Hence, homology classes do not show rapid changes while merging and forming into new classes. By contrast, normal regions have a high rate of change since cells are distinct and, when the threshold increases, the holes between them are filled rapidly. The derived profiles for tumor and non-tumor patches are distinguishable in terms of their qualitative characteristics (Fig 3).

2.2. Selection of Exemplars from CNN

The CNN used for extracting the exemplar patches for tumor and non-tumor regions is shown in Fig 4. The convolutional layers contain the set of learnt features with a combination of extremely high and extremely low frequency information. We acquire the activation matrix after CL3 by performing ReLU and MP, which rectify the feature maps by ensuring the learnt features are always positive. For the i^{th} training patch, we obtain an activation matrix $\boldsymbol{\alpha}^{(i)} \in \mathbb{R}^{H \times W \times Z}$, where H, W, and Z denote the size of the matrix in each dimension. For the architecture used in this work, $H = 24$, $W = 24$, and $Z = 108$. On the basis of the learnt features, we then calculate the mean activation value of the z^{th} learnt filter response according to (1). Let $\alpha^{(i)}_{hwz}$ denotes the (h, w, z) element of $\boldsymbol{\alpha}^{(i)}$. The mean activation

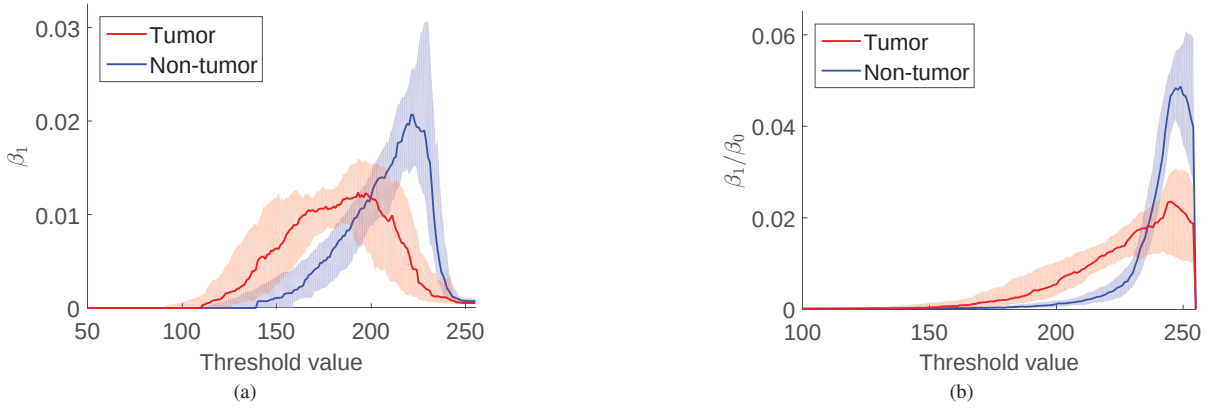


Fig. 3: Persistent homology profile for representative patches (a) β_1 and (b) β_1/β_0, showing median values with their first and third quartile, for tumor *(red)* and non-tumor *(blue)*.

value of the z^{th} filter response of $\boldsymbol{\alpha}^{(i)}$ is defined as

$$\alpha_z^{(i)} = \frac{1}{H \cdot W} \sum_{h,w} \alpha_{h,w,z}^{(i)}. \tag{1}$$

For each filter response z of the filter responses, we define

$$M(z) = \text{argmax}_i\{\alpha_z^{(i)}\} \tag{2}$$

where i runs over all patches in the training set. Then $M(z)$ is a particularly important training patch, and we use it as an exemplar. The top six exemplar patches are shown in Fig 5.

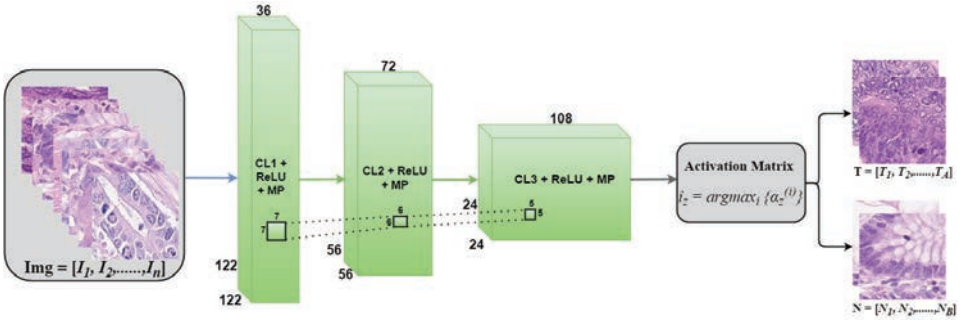
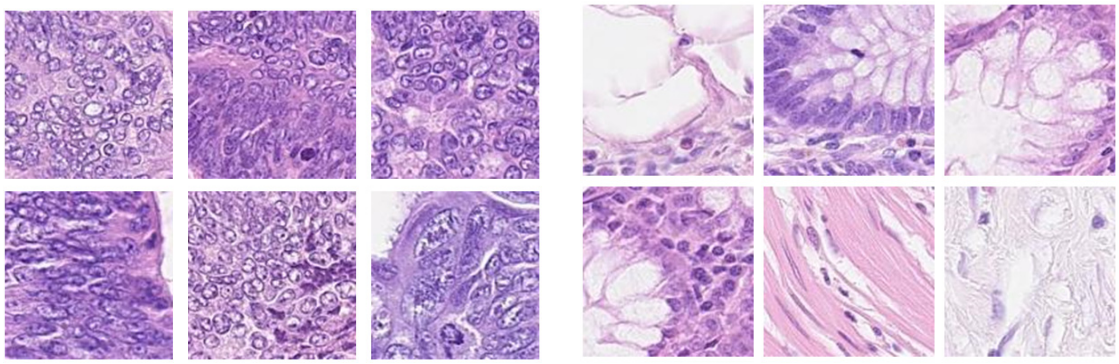


Fig. 4: Our CNN architecture has 6 layers: 3 convolutional layers, 2 fully connected layers (not shown), and 1 classification layer (not shown). An input patch is of size 128×128. The first convolutional layer (CL1) convolves the input with 36 learnt (7×7)-filters. The resulting 36-feature maps of size 122×122 are then passed through a rectified linear unit (ReLU) and max pooling (MP) is then carried out, with stride 2 in each direction. The second convolutional layer (CL2) contains 72 learnt (6×6)-filters with ReLU and an MP layer of stride 2. The CL3 layer consists of 108 learnt (5×5)-filters with ReLU and an MP layer with stride 2. After CL3, we recorded the top activations from the activation matrix. The above architecture is inspired by[8].

2.3. Patch Classification

After obtaining the homology profiles, we convert each into a discrete probability distributions, scaling the values of the functions β_0 and β_1/β_0 so that the area under each curve is one. We denote the tumor exemplar patches by $T = \{T_1, \ldots, T_A\}$, and the non-tumor exemplar patches by $N = \{N_1, \ldots, N_B\}$. We then use the symmetric Kullback-Leibler Divergence (KLD) to measure the distance between persistent homology profile of the input patch (Q) and



(a) (b)

Fig. 5: The top 6 highly activated patches from the training dataset (a) tumor (b) non-tumor.

that of each exemplar patch (P). So, associated with each input patch Q, we have computed a list of distances to the exemplar patches $D(Q) = (d_{T_1}, d_{T_2}, \ldots, d_{T_A}, d_{N_1}, d_{N_2}, \ldots, d_{N_B})$. The k nearest neighbors of Q are determined by choosing the k smallest distances in this list.

3. Experimental Results and Discussion

3.1. Dataset and Implementation Details

We collected 74 H&E stained WSIs of colorectal tumors. The tumor areas for this dataset were marked by an expert pathologist. The dataset was divided into 50 WSIs for training and 24 WSIs for testing. We randomly selected a total of 18944 (12824 training and 6120 testing) patches each of size 128×128 pixels with equal number of patches from both classes. The classification of tumor patches is fairly challenging due to the high degree of heterogeneity in tumor regions. As can be seen in Fig 6, our algorithm correctly segments the tumor region lying inside or on periphery of the hand marked tumor regions. Our expert pathologist marked the tumor regions only approximately, and so the regions marked in yellow in the "ground truth" annotations by the expert were not originally included. We chose 108 exemplar tumor patches and similarly 108 for non-tumor patches. For k-NN, we selected k=11 neighbors for classification and the value of $c(Q)$ was selected as $c(Q) = 0.2 \max_{d \in D(Q)}\{d\}$. Our algorithm was implemented in Matlab using MatConvNet with Intel Xeon E5-2887W and GeForce GTX TitanX. Finally, the segmentation accuracy was measured by the F1 score ($F_1 = 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall})$).

3.2. Results and Discussion

We compared our algorithm (PHP) with two state-of-the-art methods, including conventional CNN and HyMaP[9]. For PHP, we used variant of k-NN as a classifier (as described in Section 2.3). For CNN we used the same architecture as that described in Fig 4 with some modifications. After the convolution layers, we used two fully connected and one classification layer with the numbers of hidden neurons 512, 512, and 2, respectively. For HyMaP we used the original implementation[9] based on a set of texture features and random projections with ensemble clustering. One sees in Table 1 that PHP achieved significantly better classification accuracy than that of HyMaP[9]. PHP also achieved a slightly better precision and F1 score than conventional CNN. PHP also provided a reasonable balance between precision and recall, with the result that the number of false positives was roughly equal to the number of false negatives. Our method is computationally less expensive than either of the two currently standard techniques.

Table 1: Comparing PHP with existing state-of-the-art.

Method	Precision	Recall	F1 Score
PHP	**0.9492**	0.957	**0.9531**
conventional CNN	0.911	**0.988**	0.9524
HyMaP[9]	0.794	0.9547	0.867

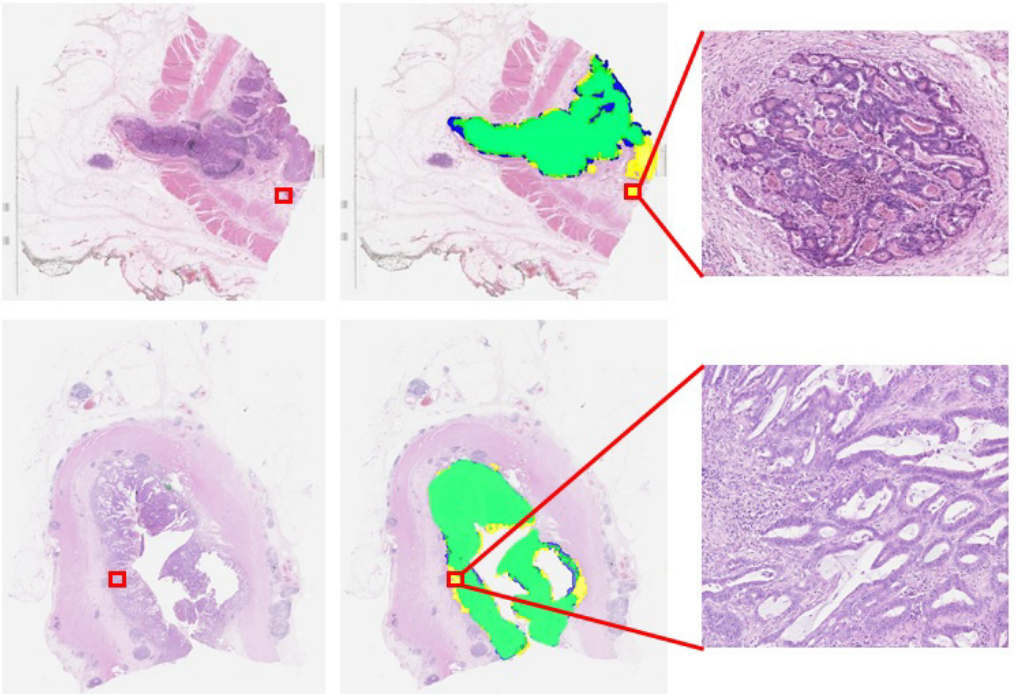


Fig. 6: Here we show the result of applying persistent homology profiles (PHP) to tumor segmentation of two different WSIs: (*Left*) Input image. (*Center*) **Green:** PHP agrees with the original expert tumor annotation. **Blue:** regions marked as non-tumor by PHP, but as tumor by expert annotation. **Yellow:** regions marked as tumor by PHP, but as non-tumor by expert annotation. (*Right*) Zoom in showing that one of our supposedly "false-positive" regions, marked by PHP as tumor and originally annotated as non-tumor, is in fact tumor.

4. Conclusions

In this work, we introduced a tumor segmentation approach for histology WSIs based on persistent homology profiles. Here we explored the degree of connectivity between nuclei using and persistent homology. We found that the tumor and non-tumor patches have distinguishable homology profiles. Experimental results using a challenging dataset demonstrate the robustness and significance of persistent homology profiles, and PHP performed better on our test dataset than standard existing methods including conventional CNN. The resulting algorithm is an order of magnitude faster then the standard CNN.

References

1. Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., Jemal, A.. Global cancer statistics, 2012. *CA: a cancer journal for clinicians* 2015;**65**(2):87–108.
2. Akbar, S., Jordan, L., Thompson, A.M., McKenna, S.J.. Tumor localization in tissue microarrays using rotation invariant superpixel pyramids. In: *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE; 2015, p. 1292–1295.
3. Al-Kofahi, Y., Lassoued, W., Lee, W., Roysam, B.. Improved automatic detection and segmentation of cell nuclei in histopathology images. *Biomedical Engineering, IEEE Transactions on* 2010;**57**(4):841–852.
4. Gurcan, M.N., Pan, T., Shimada, H., Saltz, J.. Image analysis for neuroblastoma classification: segmentation of cell nuclei. In: *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE; 2006, p. 4844–4847.
5. Edelsbrunner, H., Harer, J.. Persistent homology—a survey. *Contemporary mathematics* 2008;**453**:257–282.
6. Carlsson, G.. Topology and data. *Bulletin of the American Mathematical Society* 2009;**46**(2):255–308.
7. Khan, A.M., Rajpoot, N., Treanor, D., Magee, D.. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *Biomedical Engineering, IEEE Transactions on* 2014;**61**(6):1729–1738.
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. 2012, p. 1097–1105.
9. Khan, A.M., El-Daly, H., Simmons, E., Rajpoot, N.M., et al. Hymap: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images. *Journal of pathology informatics* 2013;**4**(2):1.